ARI Research Note 2003-06

Gesture Recognition System for Hand and Arm Signals

Donald R. Lampton
Bruce W. Knerr
U.S. Army Research Institute

Bryan R. Clark
University of Central Florida
Consortium Research Fellows Program

Glenn A. Martin
Donald A. Washburn
Carlos J. Rosas-Anderson
Institute for Simulation and Training

Simulator Systems Research Unit
Stephen L. Goldberg, Chief

November 2002

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the U.S. Total Army Personnel Command

ZITA M. SIMUTIS
Acting Director

Technical Review by
John S. Barnett, ARI

NOTICES

DISTRIBUTION: This Research Note has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This Research Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this Research Note are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)
November 2002

2. REPORT TYPE
Final

3. DATES COVERED (from...to)
July, 2001 – June, 2002

4. TITLE AND SUBTITLE
Gesture Recognition System for Hand and Arm Signals

5a. CONTRACT OR GRANT NUMBER
V1A

5b. PROGRAM ELEMENT NUMBER
2Q622785-A

6. AUTHOR(S)

Donald R. Lampton, Bruce W. Knerr, Bryan R. Clark, (U.S. Army Research Institute); Glenn A. Martin, Donald A. Washburn, and Carlos J. Rosas-Anderson (Institute for Simulation and Training)

5c. PROJECT NUMBER
A790

5d. TASK NUMBER
202A

5e. WORK UNIT NUMBER
H01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: TAPC-ARI
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Research Note 2003-06

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT *(Maximum 200 words)*:

This report describes an evaluation of a computer system for recognizing human hand and arm signals as a means of interacting with virtual environments. The system consists of two video cameras, software to track the positions of the gesturer's head and hands, and software to recognize gestures by analyzing the position and movement of the hands. The software was hosted on a standard PC. A set of 14 gestures from Army Field Manual 21-60, Visual Signals, was used to test the system. Ten participants individually performed each gesture twice as discrete trials, with a brief rest period between each trial. The average recognition rate was 68%. The highest average recognition rate for an individual was 86%; the lowest was 57%. Three of the 14 gestures were always recognized correctly, and one was never recognized correctly. While no tracking failures occurred for four of the gestures, tracking failures ranged from 10% to 100% for the other ten. The system's capabilities for untagged optical tracking and recognition of gestures involving certain types of repetitive motion advance the state-of-the-art in computer-based gesture recognition. However, for training applications, substantial improvements are needed in tracking reliability and recognition of gestures involving the depth dimension.

15. SUBJECT TERMS
gesture recognition   hand and arm signals   dismounted infantry   optical tracking

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	22	Mr. Donald R. Lampton 407/384-3989

This Page Intentionaly Left Blank

GESTURE RECOGNITION SYSTEM FOR HAND AND ARM SIGNALS

CONTENTS

List of Tables

Page

List of Figures

Introduction

This report describes an evaluation of a gesture recognition system (GRS) for recognizing hand and arm signals performed by humans. The GRS was developed by the Cybernet Systems Corporation under a Small Business Innovation Research (SBIR) Phase II contract titled, "Recognition of Computer Based Human Gestures for Device Control and Interacting with Virtual Worlds". Theoretical and technical aspects of the development and operation of the GRS are described in detail in the contractor's final report (Beach & Cohen, 2000). Aspects of the development of the system are proprietary.

The evaluation of the tracking and recognition accuracy of the Cybernet GRS was conducted within the context of a research program examining the use of immersive virtual environments for training dismounted infantry. A long-term research goal is to have a gesture recognition system that allows a trainee, immersed in a virtual environment, to control computer-generated forces by making hand and arm signals in a similar manner as a small unit leader would control a unit, for example a platoon, squad, or fire team, during real-world training exercises or actual missions.

For the short term, our goal was to measure tracking performance and gesture recognition rates under simple, controlled conditions. This approach would help identify further improvements needed in the system before it would be meaningful to test recognition within the much more complex context of a simulated infantry mission.

The background section of this report addresses the following topics: challenges of providing effective training for small unit leaders, the use of immersive virtual environments to meet training challenges, previous approaches to gesture recognition, and the Cybernet GRS.

The remainder of the report is organized as follows: the selection process is described for the 14 standard Army hand and arms signals used in the evaluation, and efforts to improve the tracking performance are listed. The characteristics, selection, and train-up of the participants who performed the gestures and the data collection procedure are described, and the tracking and gesture recognition scores for each of the 14 gestures are presented. The report concludes with a discussion of the capabilities and limitations of the current system.

The Need for Immersive Virtual Environments Training Research

The U.S. Army requires vastly improved dismounted soldier simulation capabilities to meet multiple needs. The first need is for simulations that allow dismounted leaders, soldiers and units to train effectively even if they do not have frequent opportunity to participate in high fidelity field training exercises. In addition, leaders, soldiers, and units need effective mission rehearsal tools that prepare them for specific combat operations in all types of terrain. Finally, Army decision makers need inexpensive and high fidelity prototyping and testing systems that will allow them to explore and evaluate potential doctrine, organizations, equipment, and soldier characteristics. These needs are very important today: they are likely to be critically important as the Objective Force becomes a reality.

Emerging Virtual Environment (VE) technologies, such as low cost computer image generators, locomotion platforms, intelligent computer-controlled forces, and immersive helmet mounted displays, have the potential to provide training, mission rehearsal, and experimentation capability for dismounted soldiers and leaders. However, the potential of VE is currently unrealized because the Army has not yet solved critical hardware and software limitations, documented effective training methods and strategies, or created training support packages necessary to use it.

In response to this need, the Army Research Institute's (ARI) Simulator Systems Research Unit, in conjunction with the University of Central Florida's Institute for Simulation and Training (IST), established a laboratory to conduct research on the use of immersive virtual environments for training dismounted combatants, such as infantry and special forces. One goal of the research is to develop a dismounted leader trainer at the fire team, squad, or platoon level. Leader trainees will be able to execute a series of realistic training scenarios (combat operations and support operations) in the simulator. Repeated practice, enhanced by training features, coaching, and After Action Reviews (AARs) will build decision-making and coordination skills. Computer-controlled or semi-automated agents will represent subordinates, other friendly forces, enemy forces, and civilians. The intent is to have a training system that is realistic and effective, yet requires a fairly low level of personnel support for subordinates and role players.

Such a system requires a way for these leader trainees to communicate directly with the computer-controlled entities which represent their subordinates. Automated GRS, along with automated voice recognition systems provide a way to do this.

Previous Efforts to Develop Gesture Recognition Systems

The potential for GRSs to provide alternative ways to interact with computers and to control electronic devices has inspired a considerable body of research. Several different approaches to gesture recognition are described briefly below. This short description provides a context for considering the new approach employed by the Cybernet system.

Most gesture recognition systems can be categorized as using one or a combination of the following techniques: Artificial Neural Networks. Graph Matching, Finite-State Automata, or Hidden Markov Models.

Artificial Neural Networks (ANNs) are information-processing paradigms designed to mimic the human nervous system model of information processing. With respect to pattern recognition, a one-way, feed-forward, neural network is most-often used. Feed-forward networks only allow information to flow from input to output, without feedback such as loops. The system is first trained with a set of input/output pairs. When the system is executed, it attempts to identify the input pattern and then produce the matching output "firing" pattern associated with it. Just as the human nervous system communicates through various neural firing patterns, ANNs "fire" binary codes, which indicate how similar the input pattern is to the previously trained input patterns via a threshold-type design. The unique flexibility of ANNs is in their capacity to train themselves and recognize untrained input. ANNs can observe untrained patterns, and identify the closest matching input pattern, producing output accordingly (Stergiou & Siganos, n.d.).

The Graph Matching Technique may be used for pattern recognition when data can be translated into a graphic form. By spatially transforming the graphs through methods such as rotation or scaling, the data set, which is represented by a graph, can be compared to see whether it is congruent with another set of data. Subgraph isomorphism, can be tested for, to determine whether a data set is part of another, larger data set (Bunke & Jiang, 2000).

The Finite-State Automata Technique uses the most basic functional model of a machine or computer (Hutchings, 1996). It begins with an initial state, and is presented with a string of input. The input is usually a set of events or statuses, whose relations, or transitions, to one another are coded as a finite set of symbols, referred to as the alphabet. The automaton transitions from one state to another, until it has processed all of the input and reaches a final state. The automaton then either "accepts" or "rejects" the input, based on whether or not the final state was reached by a set of transitions that were recognized by the automaton's alphabet (Daciuk, 1998).

The Hidden Markov Model (HMM) technique has been successfully employed in speech recognition applications, and is emerging as a new approach to gesture recognition. HMMs are characterized by a linear sequence of nodes, in which two states, present and new, are stochastically determined. The new state is randomly determined by the probability of certain conditions, based on the present state or node (Campbell, Becker, Azarbayejani, Bobick, & Pentland, 1996). A HMM involves an unobservable present state, which is also based on probability. Each transition involves a stochastically determined match, insert, or delete state. If it is not in a match state, the insert state may attach information, known as a residue, to the present state. If there is no residue associated with the present node, it may transition to the delete state (Karplus, 1995).

Karplus (1995) notes the benefits of using a HMM include the ability to quickly search and compute input continuously, as well as a low memory burden. A drawback to the HMM is that the transition sequences may include junk data along with the interesting data. Sometimes this can be helpful, as it may lead to the discovery of a new relationship between nodes, however it usually complicates the ability to easily interpret a sequence of nodes (Karplus, 1995).

The effort to develop gesture recognition systems has many parallels to voice recognition systems. Unfortunately, one of the commonalities is that development of highly reliable systems has proven extremely difficult.

The Cybernet Gesture Recognition System

The Cybernet Systems Corporation GRS was developed under a Small Business Innovation Research (SBIR) Phase II contract titled "Recognition of Computer-Based Human Gestures for Device Control and Interacting with Virtual Worlds". The Department of Defense (DoD) SBIR program funds early-stage Research and Development projects by small technology companies --projects that serve a DoD need and also have the potential for commercialization in private sector and/or military markets. The SBIR program is described in detail at http://www.acq.osd.mil/sadbu/sbir/. Theoretical and technical aspects of the development and operation of the GRS are described in detail in the contractor's final report (Beach & Cohen, 2000).

The Cybernet GRS hardware is composed of two Cohu 1300 series color Charge Coupled Device (CCD) cameras mounted on a precalibrated bar. Two cameras are necessary to perform three-dimensional tracking. These cameras are connected to Matrox Meteor framegrabber cards inside a Pentium III PC running a FreeBSD 4.0 operating system. Berkeley Software Distribution (BSD) is UNIX related software.

The GRS software has two separate components: tracking and recognition. Figure 1 provides a context for the description of the GRS tracking function. The figure depicts the view from the operator's station. The two images are from the two cameras. Looking at the monitor, the human operator of the system uses a mouse to place a cursor over, and click on, the image of the gesturer's left hand, right hand, and head. System software draws a box around each appendage. After this initialization process is completed, the system will track movement of the gesturer's hands. The tracking system uses the color of the hand to differentiate between the hand and the background.

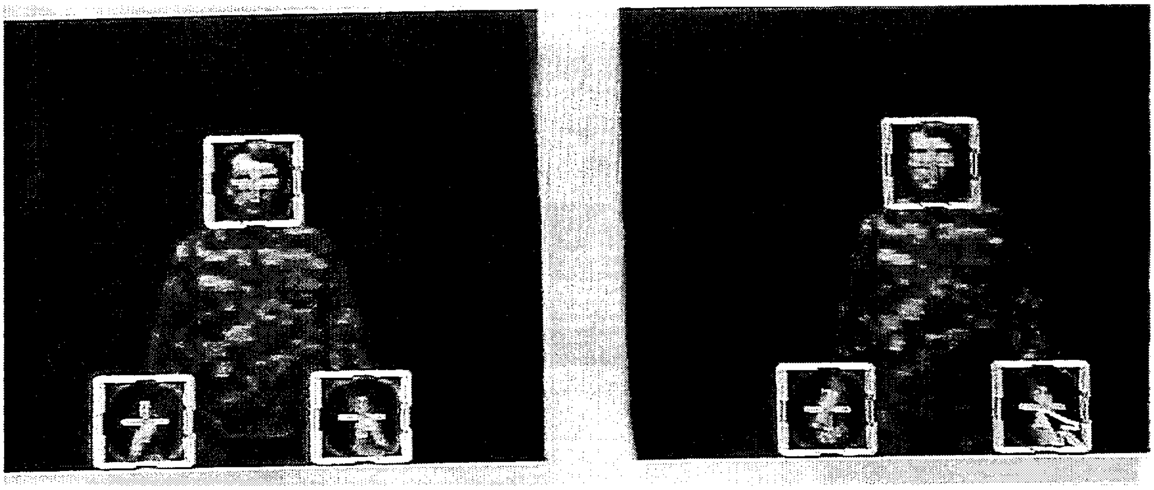

Figure 1. Initialized tracking boxes.

The tracking system feeds information about the position and movement of the hands, relative to head position, to the gesture recognition software. The gesture recognition software compares this information with definitions or examples of gestures that have already been stored in the system.

The Cybernet GRS uses two very different approaches to defining gestures, contingent upon whether the gesture involves movement. Gestures that involve movement are categorized as "dynamic"; those that do not are "static" gestures

To define a dynamic gesture, the gesturer stands before the camera system and begins making the gesture. The system operator initiates data capture mode. System software indicates when a sufficient number of data points have been captured. The operator then informs the gesturer that the gesture definition trial is over.

The mathematical analysis performed on repetitive motions is the defining factor of the Cybernet GRS. It is also the most important aspect of the system covered by the non-disclosure agreement applying to SBIR contracts.

The procedure for defining static gestures differs markedly from the dynamic procedure. The gesturer takes the pose that defines the gesture. Human data collectors use measuring devices such as tape measures to determine the distance of the hands from the head. The measurements are taken in three dimensions, corresponding to height, width, and depth. These data are then entered via keyboard to a system file.

Two concepts from voice recognition research can be used to characterize the Cybernet GRS. The term "speaker independent" describes a voice recognition system for which the system need not be retrained to recognize the voice of each unique user. The Cybernet system is designed to be "gesturer independent". The system can recognize gestures performed by individuals in addition to the individual who initially defined the gestures. This "independence" would be very valuable in most training contexts in that gestures would not have to be redefined for each trainee.

In voice recognition research, a distinction is made between continuous and discrete systems. Discrete systems require that words be said one at a time such that there is a pause between each word. Continuous systems recognize words spoken in naturally flowing phrases and sentences. The current version of the Cybernet system recognizes gestures performed one at a time.

Planned Approach

We planned to evaluate the Cybernet GRS in a research facility operated by the University of Central Florida Institute for Simulation and Training. The Cybernet GRS was to be evaluated with two different tracking systems: the untagged optical tracking system provided by Cybernet, and an electromagnetic tracking system already in use in the facility. While the optical tracking system provides a low-cost approach, it appeared to be limited to use under conditions in which the gesturer could maintain a constant orientation (i.e., facing the cameras). This might be possible in some virtual simulations or simulators, but not in others. Electromagnetic tracking, while more expensive, should also be useable when the gesturer turns freely.

We also expected that there would be two phases to the research: an initial period of try-out and formative evaluation during which system control and data collection software was developed and adjustments were made to obtain the best performance from the GRS, followed by a series of more trials.

Selection of Gestures for Evaluating the System

U.S. Army Field Manual 21-60, "Visual Signals", is a guide to commonly used Army visual signals. The stated purpose of the manual is to standardize visual signals and to serve as a training reference. The manual uses pen and ink drawings to illustrate more than one hundred visual signals. The gestures most relevant to dismounted infantry operations are in Chapter 2, "Arm-and-hand Signals for Ground Forces". Leaders of dismounted units use these arm-and-hand signals to control the movement of individuals, teams, and squads.

The gestures used in this evaluation of the GRS were selected from the perspective of using virtual environments to train dismounted infantry. We asked an Army NCO, affiliated with the ARI Fort Benning Research Unit and familiar with dismounted infantry operations in urban terrain, to identify about 10 gestures that would define a basic set of hand and arm signals that would be used frequently during infantry missions in urban terrain. We arbitrarily selected the number 10 to provide a reasonable sample of gestures with which to test the system. The NCO actually selected 12 gestures. We then asked instructors, whose responsibilities include training visual signals, to record demonstrations of the gestures on videotape. The instructors added two more gestures to the list, resulting in a videotape with 14 gestures.

The instructor who performed the gestures on the videotape used deliberate and methodical movements that clearly demonstrated the gestures. At our request, the instructor assumed a neutral position, with his hands hanging at his sides, between each gesture. An off-screen narrator announced each gesture, then the instructor performed the gesture twice.

The gestures used in the evaluation were: Attention, Ready to Move, As You Were, I Do Not Understand, Halt or Stop, Fire, Commence Firing, Increase Speed, Wedge, Line, Contact Left, Action Right, Danger Area, and Security. The gestures are depicted in Figures 2 through 14. ("Security" does not appear in the Field Manual 21-60. It requires the right hand to move in the depth dimension. Two fingers of the right hand move toward and away from the eyes, as if pointing at the eyes.)

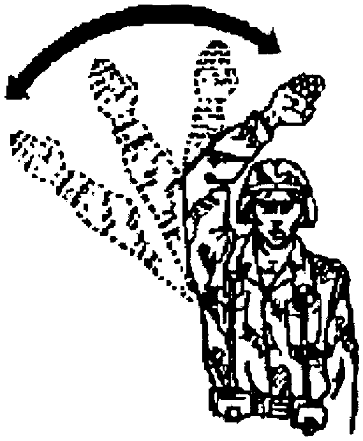

Figure 2 Attention.

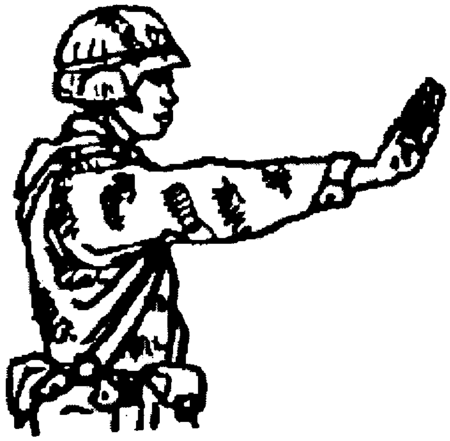

Figure 3 Ready to Move.

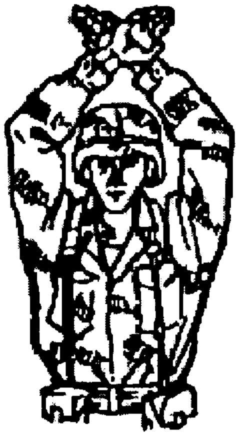

Figure 4 As You Were.

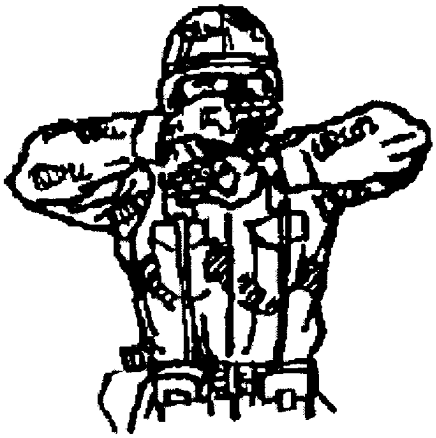

Figure 5 I Do Not Understand.

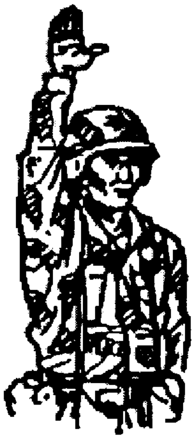

Figure 6 Halt or Stop.

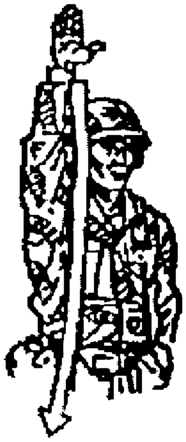

Figure 7 Fire.

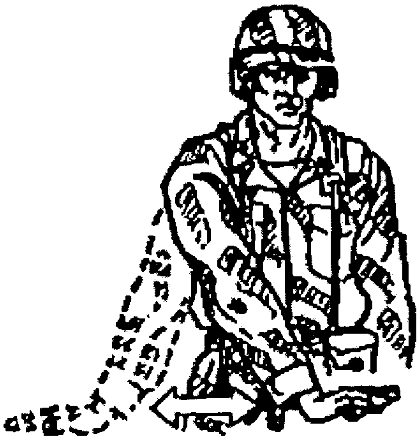

Figure 8 Commence Firing.

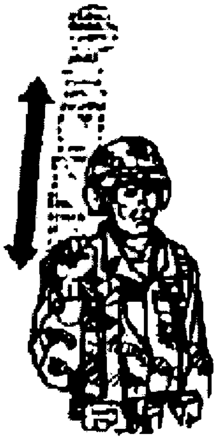

Figure 9 Increase Speed.

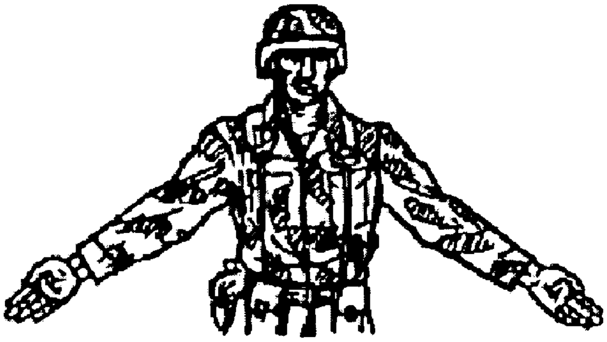

Figure 10 Wedge.

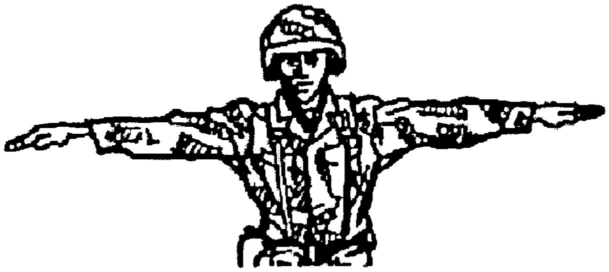

Figure 11 Line.

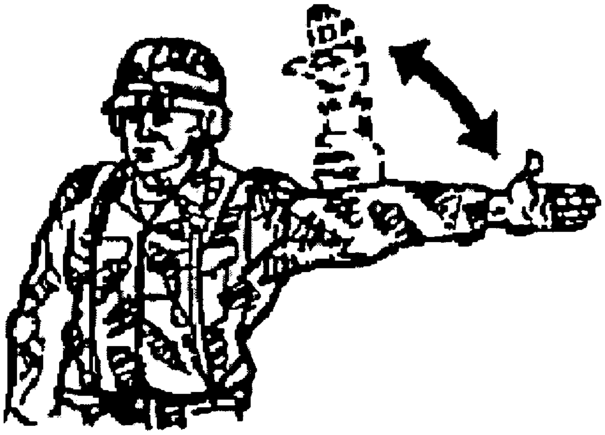

Figure 12 Contact Left.

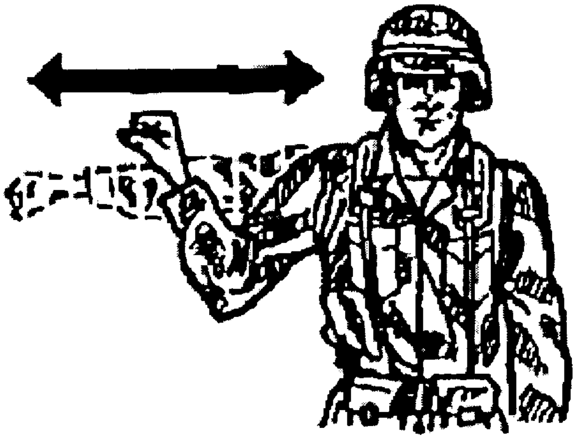

Figure 13 Action Right.

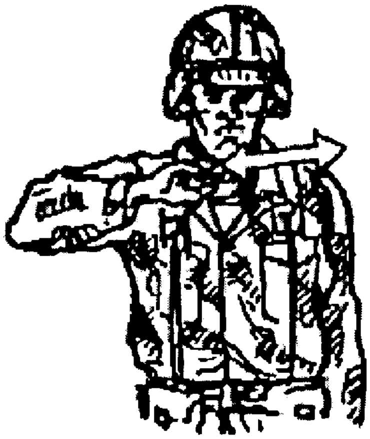

Figure 14 Danger Area.

Initial Tryouts and Formative Evaluation

When the Cybernet GRS was initially delivered there were significant problems with loss of tracking. For example, during or after the procedure in which the hands and head positions were initialized, the tracking boxes would frequently float away from the hands or head. Thus, tracking was lost even before the gesturer began to perform the gesture.

Working together, ARI and University of Central Florida Institute for Simulation and Training personnel tried several variations in room lighting, background materials, and adding optical markers for the gesturer's hands and head. Room lighting variations included using a spotlight normally used for photography sessions. We tried backgrounds of various colors and reflective properties. Gloves of different colors, wrist bands, and glow-in-the-dark light sticks were tested. Obviously, adding markers to the gesturer's hands and head is antithetical to the goal of untagged tracking. However, the initial tracking performance was so poor we were willing to sacrifice untagged tracking in order to at least be able to test the recognition software.

These manipulations failed to significantly improve tracking performance. However, demonstrating that changes in the optical properties of the test conditions were not producing corresponding changes in tracking performance aided Cybernet in identifying problems in the tracking software. Iterations of tracking testing and software fixes occurred over several months. During this testing we determined that tracking loss was more likely to occur if the gesturer wore light colored clothing or short sleeves.

In another attempt to improve tracking performance, we modified the tracking software so that the tracking box for the head was locked in place after it was initialized. Once the head tracking box was initialized it would not move even if the participant's head moved left or right. None of the gestures evaluated required any head movement. (There are at least two situations in which it would be desirable to have the capability to track head movement. The first is to allow creation of new gestures in which head movement is part of the definition of the gestures. The second is to allow the gesturer to move right or left. or forward and back. and then make a gesture. Because the head tracking box would move with the gesturer's head. and the hand tracking boxes are analyzed relative to the head tracking box. the system would still be able to recognize a gesture even though the gesturer had moved after the tracking boxes were initialized. Although these capabilities might be useful for some application. we did not need either for the current evaluation.)

In yet another effort to improve tracking performance, IST personnel attempted to integrate the Cybernet gesture recognition software with an electromagnetic tracking system. (For other applications the electromagnetic tracking system has been very precise and very reliable.) This approach was intended to bypass the problems with the Cybernet optical tracking system. However. the recognition software experienced problems involving the interpretation of z (depth) dimension data. This problem occurred with all the gestures. Beach and Cohen (2000), acknowledged a problem with the software handling of z dimension data.

IST personnel successfully created a software front-end to support data collection with the Cybernet system. The IST software addition recorded gesture recognition performance to provide a backup for manual data recording and greatly facilitated initiation of discrete data capture trials.

Evaluation of Tracking and Gesture Recognition

Participants

Ten male volunteers served as research participants. Age ranged from 18 to 52.

Procedure

The participants were run one at a time. They stood facing the midpoint between the two cameras in front of a uniform black background. Participants were provided with a Battle Dress Uniform (BDU) (see Figure 1) to wear over their clothes. The use of BDUs avoided confounding by variations in the color of the participants' clothing.

The participants were not required to memorize the gestures. They watched the videotape, on which each of the 14 gestures is demonstrated twice, to become familiar with the gestures.

The same procedure was used for recording tracking and recognition for each gesture. The videotape of a gesture was played, then the participant made the gesture; the next demonstration of that same gesture was played, and then the participant made the gesture again.

(The evaluation was to measure performance of the GRS, not to test the ability of the participants to memorize the gestures. Therefore, a data collector watched the participant perform the gesture. If either the data collector or the participant felt that the gesture had not been performed properly, then the trial was repeated. The option to repeat a trial was almost never needed.)

Each time the participant made a gesture constituted a "trial". For each trial a data collector recorded if a tracking loss occurred before, during, or after the gesture was performed. Also, the gesture recognized by the system was recorded. The GRS displayed "Null" if the gesture being performed was not matched to a stored gesture definition.

The first time a gesture was performed, the GRS operator waited until the participant initiated the gesture before activating the gesture recognition system. For lack of a better term, we refer to this as "Mid-Gesture". For the second trial for that same gesture, the operator turned on the system while the participant was in the neutral or "From Rest" position, then told the participant to begin the gesture. This is referred to as the "From Rest" condition. Thus, we measured recognition rates under two different conditions for initiating each trial: one corresponding to the way that the gestures are defined for system memory, and one corresponding to the way the gestures are demonstrated on the videotape.

When tracking losses occurred, the tracking boxes were re-initialized before starting the next trial. The gestures were always performed in the order in which they appeared on the videotape.

Results

Tracking and Recognition Rates

The Results section presents summary statistics for gesture recognition rates and tracking losses for each of the trial initiation conditions. then recognition and tracking data for each individual gesture.

As seen in Table 1, the gesture with the highest recognition rate (100%) under both conditions was Contact Left. The gesture with the worst recognition rate under both conditions was Ready To Move (0%). The average percent correct gesture recognition for both conditions was 68% although the recognition for individual gestures differed across conditions.

Averaged across the two conditions (From Rest and Mid-Gesture) the highest average recognition rate for an individual was 86% and the worst rate was 57%. The tallest (Height = 75.5"; Arm Span = 75") had an average recognition rate of 57%. The shortest (Height = 69"; Arm Span = 66") had a recognition rate of 68%.

Table 1 Percent correct recognition for each gesture

Target Gesture	Percent Correct Recognition	
	From Rest	Mid-Gesture
2-1 Attention (D)	100%	90%
2-2 Ready to Move (S)	0%	0%
2-4 As You Were (S)	80%	70%
2-5 I Do Not Understand (S)	100%	90%
2-7 Halt or Stop (S)	60%	100%
2-25 Fire (D)	30%	20%
2-26 Commence Firing (D)	100%	80%
2-33 Increase Speed (D)	80%	60%
2-37 Wedge (S)	80%	100%
2-39 Line (S)	90%	90%
2-51 Contact Left (D)	100%	100%
2-55 Action Right (D)	80%	70%
2-XX Danger Area (D)	40%	60%
2-XX Security (D)	10%	20%

Note: N=10. S = Static gesture. D = Dynamic gesture.

The average percent correct for the individual who trained the gestures was 86%, compared to an average of 66% correct for the other nine participants. The average percent correct for the individual who trained the system differed across conditions, however, as the percent correct for the From Rest condition was 79% (which three other participants scored as well), and 93% for the Mid-Gesture condition (the next highest score, 71%, was obtained by four participants).

The tracking errors can be measured from several perspectives. One approach is to record *when* a tracking error occurs. In the From Rest condition, there are three possibilities: Before, During, or After the gesture is performed. The Mid-Gesture condition only has the possibility of "During", so the two conditions cannot easily be compared on the same scale by tracking *when* errors occur. These data did provide us with a way of determining what caused an error, i.e. tracking or recognition software. Another approach to measure tracking errors is to record *where* an error occurs. One way is to note which camera loses tracking (right or left). Another, more detailed way, is to describe exactly where tracking was lost (right hand, left hand, head) per camera. These data were used to determine the presence of any errors that were equipment-specific, such as a poor camera position or damaged equipment. We did not find any such problem in our testing.

Table 2 Tracking loss

Target Gesture	Condition					
	From Rest Position				Mid-Gesture	
	Before	During	After	None	During	None
Attention (D)				100%		100%
Ready to Move (S)		10%	40%		20%	
As You Were (S)		30%	80%		60%	
I Do Not Understand (S)	10%		100%			100%
Halt or Stop (S)	30%	40%	10%		10%	
Fire (D)	20%	90%			90%	
Commence Firing (D)		10%			10%	
Increase Speed (D)		10%			10%	
Wedge (S)				100%		100%
Line (S)		10%				100%
Contact Left (D)				100%		100%
Action Right (D)				100%		100%
Danger Area (D)		20%				100%
Security (D)			40%			100%

Note: Values are the percentage of trials in which any type of tracking loss occurred. The "From Rest" condition also shows the point in time at which they occurred. N = 10. (S) = Static gesture. (D) = Dynamic gesture.

Finally, a simple approach to measuring errors is to count the number of trials in which *any* type of tracking error occurred (e.g. "Before, During, After", and right or left camera). Using this approach, out of a total of 20 trials for each gesture (10 trials per each condition), the gestures with the lowest occurrence of tracking errors (0) were Attention, Wedge, Contact Left, and Action Right, seen in Table 2 as 100% "None". The gestures with the highest occurrence of tracking errors were I Do Not Understand (100% loss "After" in From Rest condition), and Fire (90% loss "During" under both conditions) as seen in Table 2.

Furthermore, every participant performed each of the 14 gestures under two conditions, for a total of 28 trials per participant. Two participants scored the lowest occurrence of errors, with at least one tracking error in 7 of the 28 trials. The individual with the highest occurrence of errors had at least one tracking error in 13 of the 28 trials.

Analysis by Individual Gestures

Attention is a dynamic gesture in which the right arm is waved above the head. It had high recognition rates under both conditions (100% for From Rest and 90% for Mid-Gesture). This gesture is well suited to the Cybernet GRS in that it is defined by repetitive motion that occurs in a two-dimensional (x,y) plane with no depth (z) component relative to the cameras. From the tracking system perspective, Attention is a good candidate in that the palm remains exposed throughout the gesture, and the tracking boxes for left hand, right hand, and head never overlap.

Ready to Move is a static gesture. As shown in Figure 3, the gesture is defined by the right hand extending toward the observer. (The manual depicts the gesture from a side perspective because of the difficulty of representing depth in a two-dimensional drawing.) The gesture was never recognized under either condition. The Cybernet gesture recognition software almost always failed to recognize gestures that involved a non-zero depth value. Tracking loss could occur if the tracking box for the right hand overlapped with the head.

As You Were is a static gesture, with both hands crossing above the head. The recognition rates were 80% for From Rest and 70% for Mid-Gesture. This gesture is ill suited for the tracking system because of overlap of the tracking boxes.

I Do Not Understand is a static gesture, and similar to As You Were, both hands are in close contact and therefore ill-suited for the tracking system. After the gesture is recognized, tracking loss occurs as the hands are moved away from each other. The recognition rates were high:100% for From Rest and 90% for Mid-Gesture.

Halt or Stop is a static gesture in which the hand is raised high above the head. This gesture had a 100% recognition rate for the Mid-Gesture condition. The lower 60% rate for the From Rest condition reflects a loss of tracking as the hand is raised from the rest position to above the head.

Fire is the only dynamic gesture that does not involve repetitive movement. (The palm is lifted above the head and dropped once.) Lack of repetitive movement makes this a poor candidate for recognition by the Cybernet recognition software. There are also tracking problems: as the hand drops only the tips of the fingers are in sight of the cameras and tracking is lost. It is also possible that the speed (the gesture is made more rapidly than the other dynamic gestures) of the movement presents tracking problems. Fire had a low recognition rate (30% for From Rest, 20% for Mid-Gesture) and lost tracking on 90% of the trials.

Increase Speed is a dynamic gesture in which the right clenched fist repeatedly moves up and down above the head. The repetitive motion aspect of the gesture is well suited for the Cybernet GRS. However, because of the smaller area facing the cameras, the clenched fist is a more difficult tracking target than an open hand. In addition. the tracking box for the right hand passes near the tracking box for the head during the gesture. (The recognition rates were 80% for From Rest, 60% for Mid-Gesture). The GRS sometimes misidentified this gesture as "Fire".

Commence Firing is a dynamic gesture in which the right arm remains in the rest position, while the wrist moves the hand from front to side. Repetitive motion and no overlap of the tracking boxes make this a good candidate. Commence Firing had a high recognition rate (100% for From Rest, 80% for Mid-Gesture) and a low tracking loss rate, 10%.

Wedge is a static gesture in which the arms are positioned outward from the sides, at about a 45 degree angle from the body, with palms facing forward. Wedge is well suited for the tracking system in that there are widely separated tracking boxes. fully exposed palms, and no motion. Tracking was never lost. The recognition rates were 80% for From Rest and 100% for Mid-Gesture.

Line is similar to Wedge, however the arms are raised in line with the shoulders and the palms face down. The recognition rate was 90% under both conditions, and there were no tracking losses.

Contact Left is a dynamic gesture with a movement pattern very similar to Attention. The main differences are that Contact Left is performed with the left arm. and the elbow is at the same height as the shoulder - not above the head as in Attention. Contact Left was the best-recognized dynamic gesture with 100% recognition under both conditions and no tracking losses.

Action Right is a dynamic gesture similar to Increase Speed. except that the movement is horizontal. from the body to away from the body, again with a clenched fist. and "punching" like motion. Unlike Increase Speed. Action Right does not require the tracking box for the hand to approach and possibly overlap the tracking box for the head. Action Right had a recognition rate of 80% for From Rest and 70% for Mid-Gesture. There were no tracking losses.

Danger Area is a dynamic gesture in which the right hand moves in an diagonal throat-cutting motion. Because the tracking box for the right hand passes near the head this is yet another gesture that presents a problem to the tracking system. The recognition rates were 40% for From Rest and 60% for Mid-Gesture.

Security requires the right hand to move in the depth dimension. Two fingers of the right hand move toward and away from the eyes. as if pointing at the eyes. The recognition rates were

10% for From Rest and 20% for Mid-Gesture. These recognition rates are low. However, it was not expected that there would be any correct recognition because the gesture is defined by movement along the z (depth) axis and the movement is over a very short range. It may be that the system responded to the gesture on the basis of some feature other than the back and forth movement of the fingers. The tracking loss was not as high as expected given that by definition the right tracking box is over the head tracking box.

Discussion

In general, the gesture set was not a good match for the Cybernet system. Many of the gestures were problematic in terms of tracking (overlapping tracking boxes), recognition (lack of repetitive movement), or both.

The gestures were selected according to their relevance to urban missions, not for their suitability for use with the Cybernet GRS. The Cybernet approach would be more useful for a new application in which the gestures could be selected or created. For example, it might be useful to be able to control robotic vehicles via hand and arm gestures. If there were not a requirement that the gestures correspond to some existing system of gestures, then control gestures could be selected based on their compatibility with the Cybernet system. That is, the gestures would be defined by repetitive motions that did not require the tracking boxes to approach each other at any time during the gesture.

It is also possible that the Cybernet approach for recognizing dynamic gestures could be incorporated in a hybrid approach using other gesture recognition techniques to handle the types of gestures ill-suited for the Cybernet system. Thus, in the long term, the Cybernet approach to recognizing certain types of dynamic gestures may significantly advance the state-of-the-art of gesture recognition. In contrast, the approach to recognizing static gestures seems simple and perhaps simplistic.

The concept of untagged tracking is attractive from the perspectives of convenience and cost. However, the high tracking failure rate obtained clearly indicates that this aspect of the technology requires improvement. Our evaluation was conducted under nearly ideal conditions: good and constant lighting, close attention to proper positioning of participants prior to the gesture, and controls to insure that the gestures were indeed performed properly. We expect that performance under the conditions under which we wish to use the tracking system would be much worse. The optical tracking system was unable to track gestures in which the hands crossed the face or each other, gestures in which front to back movement was critical, gestures in which a hand changed shape, or rapid movements. Electromagnetic tracking may have performed better, and resulted in better recognition performance, but software problems prevented conduct of that part of the evaluation.

We chose to use only male participants in this experiment because our target users, close combat Infantry soldiers, are exclusively male. Although our sample included a wide range of heights and arm spans, inclusion of female participants would likely have increased the variability of the sample along those dimensions. It might or might not have made recognition more difficult.

Because of the high rate of tracking loss, the system is not currently suitable for actual training applications. Tracking loss would be especially problematic during a mission exercise because tracking loss not only interferes with gesture recognition but also requires re-initialization of the tracking boxes.

The minimum acceptable recognition rate for training applications has not been established. It could be argued that in live training, and in combat, hand and arm signals are not always recognized perfectly. Therefore, less than perfect recognition in a training simulation might be acceptable because it would reinforce the point that after the leader gives a hand and arm signal, the leader must then monitor that the rest of the squad has correctly perceived and responded to the signal. Ultimately, this question will require empirical testing. It is likely the acceptability will not depend solely on an average recognition rate, but also on the pattern of errors. In general, a failure to recognize a gesture would be less disruptive than a misperception. Failure to recognize a gesture would merely require the leader to make the gesture again. Misperception of a gesture would require the leader to stop the action resulting from the misperception, and then repeat the gesture. The degree to which an exercise was disrupted would depend on the particular misconception. If the recognition system misperceived the "line" hand and arm signal instead of 'wedge", that would not be especially disruptive. However, misperception of "fire" instead of "halt" would greatly compromise the effectiveness of a training exercise.

References

Beach. G. & Cohen, C. J. 2000. *Recognition of computer-based human gestures for device control and interacting with virtual worlds.* (Final Report prepared under contract DASW01-99-C-0004). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bunke, H. & Jiang, X. (2000). Graph matching and similarity. Teodorescu, H. -N. et al. (Eds.), *Intelligent Systems and Interfaces* (pp. 291-304). Netherlands: Kluwer Academic Publishers.

Campbell, L., Becker, D,. Azarbayejani, A., Bobick, A., & Pentland, A. (1996). Invariant features for 3-D gesture recognition. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition* (pp. 157-162). Los Alamitos, California: IEEE Computer Society Press.

Daciuk, J. (1998). *Finite state automata.* Retrieved May 31, 2001, from http://www.eti.pg.gda.pl/~jandac/thesis/node12.html

Hutchings, D. (1996). *Pattern Matching: CS 280 Lab.* Retrieved May 31, 2001, from http://www.cs.oberlin.edu/classes/dragn/labs/patmatch/patmatch20.html

Karplus, K. (1995). *Using simple Markov models to search DNA databases.* Retrieved June 8, 2001, from http://www.cse.ucsc.edu/~karplus/ismb95-submit/ismb95-submit.html

Stergiou, C. & Siganos, D. (n.d.). *Neural networks.* Retrieved June 4, 2001, from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

United States Army (1987). *Field Manual 21-60, Visual Signals.* Washington, DC: Author.